离群专利视角下的新兴技术预测*
——基于 BERT 模型和深度神经网络

■ 孔德婧[1]　董放[2]　陈子婧[3]　刘宇涵[3]　周源[2]
[1] 北京邮电大学现代邮政学院　北京 100876　[2]清华大学公共管理学院　北京 100084
[3] 华中科技大学机械科学与工程学院　武汉 430074

摘　要：[目的/意义]由于新兴技术本身的超前性，其刚出现的关注度往往不是很高。目前研究更多遵循技术发展路径依赖进行新兴技术的识别，会忽略一些颠覆现有技术轨道的技术研发。通过对与领域内主流技术相似度较低的离群专利进行分析，可以更有效地识别这类技术研发并预测新兴技术。[方法/过程]提出一种基于深度学习的离群专利识别与新兴技术预测方法。首先使用 BERT 预训练模型基于专利文本构建相似度网络，识别离群专利，然后基于 DNN 模型构建离群专利指标与技术影响力之间的关系，实现从海量离群专利中快速、准确地预测新兴技术。最后以数控系统领域为例，从德温特专利数据库获取近 10 年领域内所有专利，进行实证分析。[结果/结论]数控系统领域的实证分析结果验证了模型的有效性，同时对国家的技术发展政策制定以及相关领域企业技术布局具有重要的指导意义。

关键词：新兴技术　深度学习　离群专利　数控系统
分类号：G250
DOI：10.13266/j.issn.0252-3116.2021.17.013

1 引言

十九届五中全会指出，加快发展现代产业体系，推动经济体系优化升级，要发展战略性新兴产业。新兴技术作为一种全新的“突破性”技术，是战略性新兴产业的重要支撑[1]。其有可能颠覆现有的技术体系和原有的技术范式，使现有的产品、工艺或服务具有前所未有的性能，或者实现现有性能的大幅提高并降低成本[2]。新兴技术对市场规则、竞争态势、产业边界具有决定性的影响，甚至可能引起产业的重新洗牌[3]。因此，新兴技术预测对国家、企业等各个层面的技术布局和战略制定具有重要意义。

新兴技术在短期内快速发展，具有高度不确定性，在未来极有可能推动技术进步，且具有较大社会影响力[4]。现有大部分研究以技术发展路径依赖为基础识别新兴技术，关注于主流技术轨道中的热点、前沿技术。由于其超前性，新兴技术很可能脱离原有的主流技术轨道，并且在短时间内难以完成技术转化，而在未来会对行业发展作出极大贡献[5]。比如中国科学院过程工程研究所于 1998 年获得授权的一项可降低废气排放的解耦燃烧技术[6]，当时人们尚未意识到氮氧化物排放的危害，导致该专利成果于 2017 年才得到大面积推广应用。日本佳能公司于 1982 年申请打印机液体喷射记录头专利[7]，然而当时喷墨打印机尚未进入主流市场，直到 1990 年该专利才开始出现大量引用，引用者不乏惠普、施乐、谷歌等大企业。采用路径依赖的方法识别新兴技术则不容易及时发现这些颠覆原有技术轨道的研发。新兴技术出现早期的最基本特征是激进的创新性[8]，其形成初期往往与主流技术范式具有较低的相似度和关联性[9]，从而呈现出一种“离群”状态。从“离群点”的视角出发，能够更准确地反映新兴技术在形成初期的状态。

* 本文系国家自然科学基金项目“基于多源知识图谱的产业融合路径及机制研究”(项目编号:72004016)和国家自然科学基金项目“基于多源异构网络视角的新兴产业创新扩散作用机制及政策研究”(项目编号:71974107)研究成果之一。
作者简介：孔德婧(ORCID:0000-0002-2575-3514)，讲师，博士；董放(ORCID:0000-0003-4271-9702)，博士研究生；陈子婧(ORCID:0000-0001-7761-5810)，硕士研究生；刘宇涵(ORCID:0000-0002-3574-8479)，硕士研究生；周源(ORCID:0000-0002-9198-6586)，副教授，博士，博士生导师，通讯作者，E-mail:zhou_yuan@mail.tsinghua.edu.cn。
收稿日期:2020-12-21　修回日期:2021-05-31　本文起止页码:131-141　本文责任编辑:徐健

目前,基于"离群点"视角的新兴技术识别研究较少,已有的研究主要采用专利数据通过识别离群专利来预测新兴技术[10-11]。这些研究从"离群点"的视角做出了有价值的探索,然而仍存在一些局限性:①使用引文耦合的方法计算专利的相似度,从而得到与主流技术相似度较低的离群专利[12],缺乏对专利文本语义信息的理解,计算的相似度不够准确;②基于专利指标和专家判断的预测方法[13]成本高、耗时长。数据驱动的深度学习方法可以在保证预测效果的同时,大幅提高预测效率,降低预测成本[14],实现从海量专利中快速、准确地识别新兴技术。

从"离群点"的视角,笔者采用深度学习方法构建基于词向量和深度神经网络(deep neural networks,DNN)的新兴技术预测模型。首先使用 BERT 预训练模型将专利文本向量化,基于语义相似度构建专利相似度网络,识别出网络中的离群点作为备选新兴技术;然后,使用 DNN 模型学习离群专利的各项指标与技术影响力大小之间的关联关系;最后,利用该关系模型预测当前年份的离群专利未来的技术影响力,发掘在当前未被关注而在未来可能产生巨大影响的离群专利,从而预测新兴技术。与此同时,本研究以数控系统领域为实证案例,验证方法的有效性。

2 相关研究

2.1 新兴技术预测方法

传统的新兴技术预测主要依靠专家知识,比如德尔菲法[15]、层次分析法(Analytic Hierarchy Process,AHP)[16],在利用这些方法进行新兴技术预测的过程中,提出了许多用于描述新兴技术特征的专利指标[17-18]。其中一部分指标不会随时间的推移而改变,如 IPC 数量[19]、发明人数量[20]、非专利文献引用[21]等。还有一部分指标随时间推移会发生变化,如前向引用[22]、专利修改次数[23]等。然而仅采用这些方法难以预测复杂的技术增长与应用扩张[12]。近年来,计算能力的提升使得"数据驱动"成为可能,与此同时,随着人工智能技术的发展,基于机器学习和深度学习的新兴技术预测方法引起了广泛关注[24]。D. Kong 等以工业机器人领域为例,使用结合专家知识的机器学习方法识别高质量专利,分析技术创新缺口[25];周源等以生物信息领域为例,结合引用网络聚类与隐含狄利克雷分布(Latent Dirichlet Allocation,LDA)模型识别新兴技术领域融合演化过程。相比机器学习,深度学习具有更复杂的模型结构,模型效果更好[26]。S. Hassan 等使用包含 64 维指标的样本数据进行引文重要性预测,发现深度学习模型对高维指标的预测效果比机器学习模型更好[27];Y. Zhou 等针对专利数据量有限的问题,提出一种基于数据增强与深度学习的新兴技术预测方法[28]。

专利是识别新兴技术的重要数据来源[12]。基于专利数据的新兴技术识别研究对技术的定义分为两类:①从技术角度出发,将"一项技术"定义为属于同一个 IPC 或者使用聚类方法划分到同一簇团的所有专利。Y. Geum 等通过分析属于各 IPC 的专利特征来预测新兴技术[29];G. Kim 和 J. Bae 将专利文本聚类,然后分析每个簇团中专利的前向引用、同族专利、独立要求等指标以识别新兴技术[30];Y. Zhou 等提出一种半监督主题聚类模型,以 3D 打印领域为例,通过对簇团形成句子级的语义描述识别新兴技术[31]。②从专利的角度出发,把一项专利看作一个理论焦点(theoretical focal point),旨在通过识别高影响力专利发现新兴技术。侯建华和朱晓清以固体氧化物燃料电池技术为例,从技术发展趋势、技术成熟度和演化方向 3 个方面构建专利指标,通过 CiteSpace 中的 Sigma 指标进行技术预测[32];C. Lee 等结合多种专利指标,使用机器学习方法评估专利价值,从而预测新兴技术[33]。

笔者采用对技术的第二类定义,认为每项专利代表一项技术研究焦点。使用文献计量方法从专利中提取技术特征,使用深度学习模型预测其未来发展成为新兴技术的可能性。

2.2 离群专利识别

新兴技术的一个重要特征是激进的创新性[8],这意味着新兴技术较大可能与已有技术具有极强异质性。基于此,有学者指出识别新兴技术的过程应该更加关注"离群专利"[10-12]。离群专利所代表的技术极有可能引起技术范式的转变[11]。B. S. Aharonson 和 M. A. Schilling 认为离群专利相较于其他专利更有可能发展成为新兴技术,在专利分析过程中舍去离群专利将造成严重的信息丢失[10]。

现有研究中,离群专利的识别主要基于两种方法:引文耦合与文本相似度。K. Song 等认为拥有更多共引关系的专利相似度更高,提出一种基于专利引文耦合的方法识别离群专利[12];曹艺文等则认为专利引用本身会回避相似专利以避免对自身的创新性造成威胁,仅使用专利引用信息判断专利相似度具有一定局限性[34];J. Yoon 和 K. Kim 提出一种基于 SAO 的语义向量计算方法,基于专利的语义相似度得到离群专

利[11]。基于语义的相似度计算方法更加精确，然而由于 SAO 存在运算效率低、语义混淆等问题难以在大规模数据中应用[10]。Y. Zhang 等使用词向量方法从大量文本中提取潜在语义信息，发现深度学习在大规模文本信息提取任务中具有较好表现[35]；J. Devlin 等提出 BERT 模型，能够大幅提升现有语义表示方法的性能[36]。笔者将使用 BERT 模型提取专利文本信息，基于专利的文本相似度识别离群专利。

综上所述，从离群专利的视角对新兴技术进行预测是一种有效途径，能够较早地发掘不易被察觉的技术点，而深度学习方法在文本信息提取以及预测任务中都具有优越性能。笔者将在此基础上从离群专利的视角利用深度学习方法对新兴技术进行预测。

3 研究方法

图 1 展示了笔者提出方法的整体流程，共分为 5 个主要步骤：①专利数据获取；②使用词向量模型将专利数据进行文本向量化，根据文本相似度构建专利相似度网络，筛选出离群专利作为备选新兴技术；③使用文献计量方法从专利数据提取能体现备选技术早期特征的各项指标，并评估备选技术未来的技术影响力；④使用深度学习模型拟合专利指标与未来技术影响力之间的关系；⑤模型性能评估。

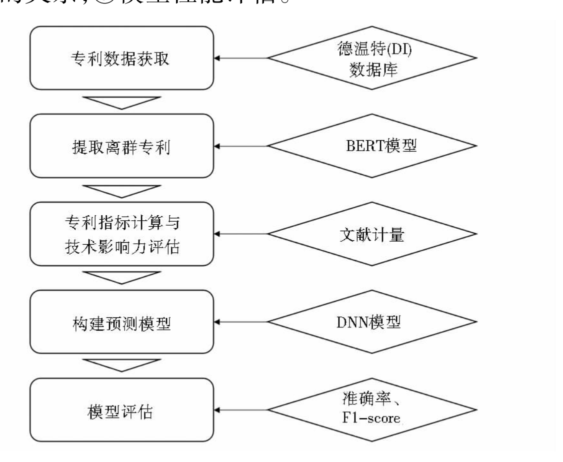


图 1　方法整体流程

3.1　获取离群专利

从专利数据库中获取到目标领域内的所有专利后，需要采取一定策略来识别离群专利。如图 2 所示，在专利相似度网络中，一篇专利为一个节点，节点之间是否有连线取决于两个专利节点的相似度。离群专利即专利相似度网络中与其他专利节点都没有连接的“离群点”。

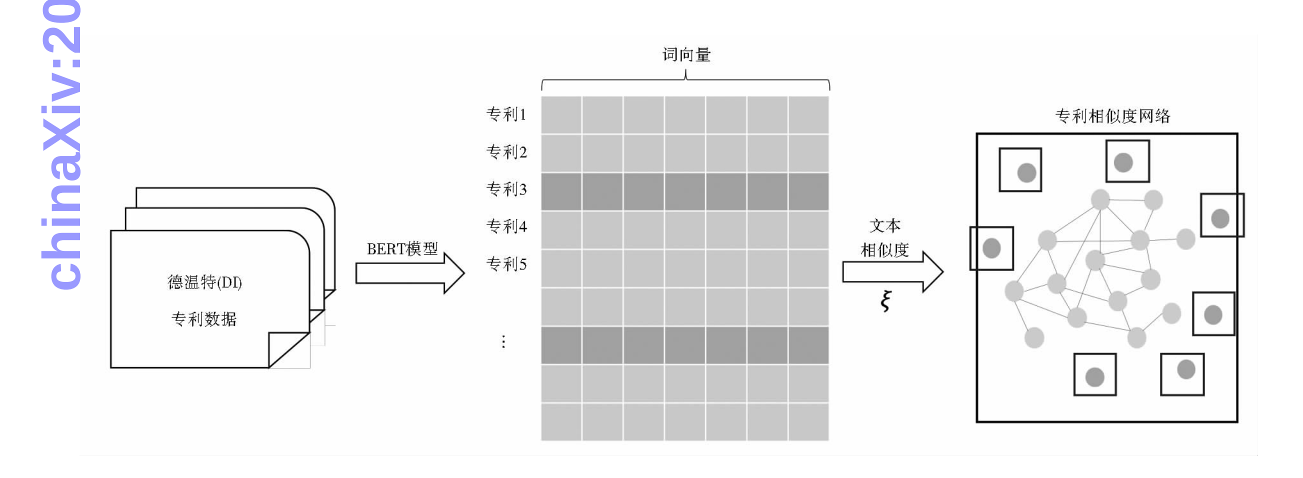


图 2　离群专利获取

离群专利的获取分为以下 4 步：①对下载的每篇专利使用 BERT 预训练模型转化为 n 维向量表示；②计算专利两两之间的相似度；③以专利为节点，相似度高于阈值 ξ 的专利间形成连线，构建专利相似度网络；④筛选出专利相似度网络中没有连线的节点，即为离群专利。

将文本向量化的过程可以称之为编码，而 BERT 预训练模型[36－37]是一种有效的基于深度学习的编码器，相较传统基于词频的方法，BERT 模型还考虑了单词间的关联关系，因此能够提取更全面的文本信息。BERT 预训练模型的输入为专利文本，每一项文本数据会被拆解成 3 个部分，如公式(1)所示：

$$X_{emb} = T_{emb} + S_{emb} + P_{emb} \qquad \text{公式(1)}$$

第 1 部分是 token embedding，携带有文本中词语自身的含义；第 2 部分是 segment embedding，用于表征长文本中句子与句子之间的上下文关系；第 3 部分是 position embedding，用于表征词语之间的顺序关系。这 3 个部分进行组合后共同作为 BERT 模型的输入，用于

编码。

由于在文本信息分析过程中,不同词语的重要性可能不同,所以在编码过程中,BERT 引入了多头自注意力机制(multihead self-attention mechanism),提高一些关键词语在分析过程中的权重,进而提高编码的准确性。首先对输入 X_{emb} 进行线性变化,得到 3 个矩阵 Q、K、V,分别为每个单词的查询向量、键向量以及值向量构成的矩阵。具体实施过程通过定义 3 个矩阵 W_Q、W_K、W_V,与输入 X_{emb} 相乘得到,即 $Q = X_{emb} W_Q$,$K = X_{emb} W_K$,$V = X_{emb} W_V$。

自注意力机制的实施过程如公式(2)所示:

$$attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad 公式(2)$$

其中,QK^T 表示当前单词与句子其他部分的关联程度,d_k 为 Q、K 的向量维数。softmax 后的值即文本中每个位置单词的权重,与值向量矩阵 V 相乘即得到一个文本中所有单词加权后的向量表示。

多头注意力机制即为同时实施多个注意力机制,将多个注意力机制的结果进行合并,完成编码。过程如公式(3)所示:

$$Y_{emb} = \text{feed}(W_z \cdot multihead\ attention(Q,K,V) + X_{emb}) + X_{emb} \quad 公式(3)$$

其中,Y_{emb} 为编码结果,feed(x) 为前向传播过程,$W_z \cdot multihead\ attention(Q,K,V)$ 表示将多头注意力机制结果经过全连接层进行合并,W_z 为全连接层的权重矩阵。

上述过程完成了一次编码,通过多次编码(将 Y_{emb} 再次带入 X_{emb})可以得到最终的编码结果,即专利文本中每一个词映射为 n 维向量,对文本中每一个词的编码结果进行加和,得到专利文本的向量表示。

专利相似度使用余弦相似度进行计算。对于两个 n 维专利向量 $x = (x_1, x_2, \cdots, x_n)$、$y = (y_1, y_2, \cdots, y_n)$,计算方法如公式(4)所示:

$$cos(\theta) = \frac{\sum_{i=1}^{n}(x_i y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad 公式(4)$$

3.2 离群专利指标提取与技术影响力评估

3.2.1 离群专利指标提取

在文献计量领域有很多描述新兴技术的专利指标。如表 1 所示,本研究使用 5 类共 11 项指标对离群专利各方面特征进行测度。

(1)新颖性。新颖性由技术创新性(technological originality,TO)以及先验知识量(prior knowledge,PK)表示。技术创新性描述专利参考其他技术领域的多样性程度,专利越广泛地结合不同领域的技术思想,就越可能产生高价值的技术发明[38-39]。先验知识描述专利对其他技术的参考程度,专利引用其他专利越多,其新颖性和商业价值就越低[40]。

表 1 专利指标及描述

维度	指标	描述
新颖性	技术创新性(TO)	引用专利的领域集中度
	先验知识量(PK)	后向引用次数
发展速度	技术生命周期(TCT)	引用专利平均年龄
知识密度	科学知识量(SK)	非专利文献引用
应用范围	技术范围(TS)	专利所属类别数
	商业范围(CS)	同族专利数量
	独立权利要求(PCID)	独立权利要求数
	从属权利要求(PCD)	从属权利要求数
发展能力	专利权人合作程度(COL)	多个专利权人则为 1,否则为 0
	发明人数量(INV)	发明人数量
	专利权人能力(TKH)	专利权人发表总专利数

(2)发展速度。发展速度由技术生命周期(technology cycle time,TCT)表示。技术生命周期能够表征技术先验知识的新旧程度以及发展快慢,从而反映出技术的发展速度[41-42]。

(3)知识密度。知识密度由专利中的科学知识量(scientific knowledge,SK)表示。专利中的知识密度越大,就越可能带来创新性、高影响力的技术发明[43-44]。

(4)应用范围。应用范围由技术范围(technological scope,TS)、商业范围(commercial scope,CS)以及专利保护范围中的独立权利要求(protection coverage described in independent claims,PCID)、从属权利要求(Protection coverage described in dependent claims,PCD)表示。技术范围表征专利在技术领域的覆盖范围。有研究表示专利所属的专利族规模越大,越可能具有高商业价值[45],因此专利中的同族专利信息可测度专利在商业层面的应用范围。独立权利要求与从属权利要求体现了专利受保护的范围。

(5)发展能力。发展能力由专利权人合作程度(collaboration,COL)、发明人数量(inventors,INV)和专利权人能力(total know-how,TKH)表示。专利权人的合作对专利价值有显著积极的影响[46-47],多发明人的专利同样具有更高价值[46]。专利权人的能力水平将会影响技术的未来发展以及影响力。

3.2.2 离群专利的技术影响力分类标签

专利的前向引用数是使用最广泛的技术影响力评估方法,它反映出专利所代表的技术对之后技术发展

的贡献程度。一项技术被越频繁、越广泛地应用到未来技术之中，意味着它具有越大的技术影响力[29,48]。因此当前年份下载的专利数据中，专利的前向引用次数可以表征自该专利被发表以来，到目前为止的技术影响力。笔者将技术影响力分为高、低两个等级，根据样本前向引用次数分布情况设定一个临界值，认为引用次数低于临界值的专利为低技术影响力样本 TL0，高于临界值的专利为高技术影响力样本 TL1。

3.3 基于离群专利的新兴技术预测

使用离群专利预测新兴技术的关键是构建出离群专利指标与其未来影响力之间的关系模型。由于一个技术领域往往会有大量离群专利，且描述备选新兴技术的专利指标较多，因此笔者使用深度学习模型 DNN 来拟合专利指标与技术影响力之间的关系。将专利数据集分为如图 3(1)所示两部分：使用过去 5－10 年的专利数据作为数据集，评估技术在当前的影响力，构建专利指标－技术影响力关系模型；再使用构建的关系模型，使用近 5 年的专利数据预测当前技术未来的技术影响力。

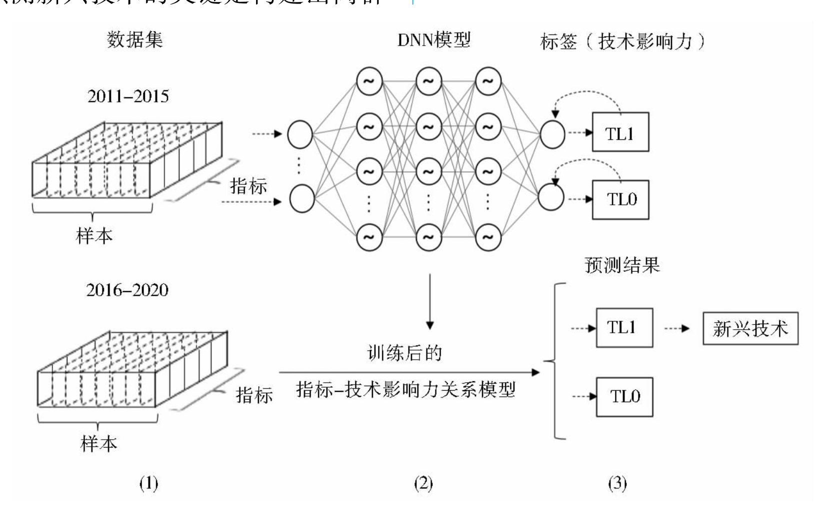


图 3　基于专利指标－技术影响力的新兴技术预测模型

如图 3 所示，DNN 模型由输入层、隐藏层以及输出层构成。模型的输入为离群专利的指标向量，输出为预测的技术影响力分类，隐藏层的激活函数均使用 Relu。数据集中每个样本均为 12 维向量，其中前 11 维为根据 3.2.1 节方法计算的离群专利指标，最后一维为根据 3.2.2 节方法得到的技术影响力分类标签。

DNN 模型的构建分为训练和测试两个步骤：

在模型训练过程中，各隐藏层参数将初始化为符合正态分布的随机值，输入训练集样本，然后将每个训练样本的预测结果与实际结果标签进行对比，使用交叉熵计算损失函数以更新各隐藏层的参数，当损失函数收敛到一定值则模型参数优化完成，停止训练。

模型测试过程使用测试集评估模型在未知样本中的预测效果。评估模型性能的指标使用准确率(accuracy)、精确率(precision)、召回率(recall)以及 F1-score。4 个指标的计算方法如公式(5)－(8)。其中，准确率评估模型预测结果整体的正确性。精确率评估模型的查准率，召回率评估模型的查全率。F1-score 综合考虑了精确率以及召回率，是两者的调和平均数，可以体现模型对不同分布情况样本的预测效果，是分类问题中衡量模型整体性能的常用指标。

$$Accuracy = \frac{TP + TN}{P + N} \qquad 公式(5)$$

$$Recall = \frac{TP}{TP + FN} \qquad 公式(6)$$

$$Precision = \frac{TP}{TP + FP} \qquad 公式(7)$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad 公式(8)$$

在上述公式中，P 为正样本，N 为负样本，TP、TN 分别为为判断正确的正样本和负样本，FP、FN 分别为判断错误的正样本和负样本。

4 研究结果

4.1 数据收集与离群专利识别

笔者选择数控系统技术领域验证方法的有效性。数控系统作为实现制造业企业数字化转型，国家智能制造战略目标的关键技术，预测该领域未来可能出现的新兴技术，对企业和国家优化技术布局、把握发展机遇有着重要意义。

本次实验从德温特(Derwent Innovation，DI)数据

库根据制定的数控系统领域检索式获取了该领域中优先权年在 2011 年 1 月 1 日至 2020 年 10 月 30 日共 58 021条专利数据。其中 2011－2015 年的22 418条专利数据用于构建专利指标－技术影响力关系模型，2016－2020 年的 35 603 条专利用于对 2025 年数控系统领域新兴技术的预测。

离群专利筛选过程首先使用 python 的 transformers 库中的 BERT 模块对专利文本进行编码，模型包含两个编码层，多头自注意力机制头数为 8 头，最终将每条专利文本映射为 128 维向量，并使计算两两专利向量的余弦相似度。接下来，将 min-max 归一化后的文本相似度高于阈值的专利之间形成连接，分别构建 2011－2015 和 2016－2020 两个专利相似度网络，并分别筛选出两个时间段内的离群专利。现有的离群专利相关研究主要通过实验的方法选取合适的相似度阈值[11－12]。当相似度阈值为 0.5 时，2011－2015 年离群专利过少，这可能会遗漏大量的颠覆现有技术轨道的技术研发的相关专利；当阈值为 0.7 时，2011－2015 年离群专利过多，这可能会引入大量干扰，降低从离群专利中识别新兴技术的效率。因此，本文最终确定阈值为 0.6，其中 2011－2015 年共有离群专利 2 747 篇，2016－2020 年共有离群专利 15 385 篇。这些离群专利分别为 2020 年、2025 年的备选新兴技术。

4.2 指标提取与模型训练

4.2.1 指标提取

按照 2.2.1 中的方法对 2020 年的备选新兴技术提取专利指标以及技术影响力，形成用于构建关系模型的样本数据集。数据集共有 2 747 个样本，其中每一个样本表示一个备选新兴技术，包含 12 维数据，其中前 11 维为专利指标，最后一维为技术影响力标签。

备选新兴技术在各专利指标及技术影响力上的表现情况如表 2 所示。整体来看，样本在各项指标上的数值跨度均较大，因此在生成 4.2.2 中使用的数据集时对所有专利指标取对数处理，负值取零。根据3.2.2 节，笔者将专利划分为高技术影响力和低技术影响力两类，其中技术影响力 TL 的衡量方法是专利的前向引用次数。因此需要对专利的前向引用次数值取一个临界值以划分高技术影响力和低技术影响力。高技术影响力专利的前向引用大于 10 是比较合理的值[33]。笔者采取 3σ 准则来确定准确的前向引用数量临界值：首先计算专利前向引用次数的均值 μ 和标准差 σ，当临界值为均值右偏一个 σ 时，其值为 16（$\mu+\sigma=16$），即当专利的前向引用次数大于等于 16 时为技术影响力为高，小于 16 时技术影响力为低。

表 2 各专利指标描述性统计

变量	观察值数量	均值	标准差	最小值	最大值
to	2 747	1.300 692	6.850 224	0	173
pk	2 747	9.942 119	64.421 84	0	2 724
tct	2 747	6.633 964	9.511 655	－0.7	257.1
sk	2 747	3.195 486	24.246 04	0	1 092
ts	2 747	2.746 269	2.882 258	1	33
cs	2 747	2.310 885	3.212 123	1	102
pcid	2 747	1.810 339	2.190 839	1	55
pcd	2 747	7.825 992	7.803 692	0	80
col	2 747	1.069 894	0.379 068 3	1	13
inv	2 747	3.596 287	2.567 181	0	20
tkh	2 747	148.772 5	286.590 5	1	1 659
tl	2 747	5.5132 87	9.559 799	0	158

4.2.2 模型训练

实际中，尽管每年均有大量技术专利产出，只有少量专利能够在未来获得较高的技术影响力，成为新兴技术。在本文 4.2.1 中得到的离群专利中，高技术影响力专利数与低技术影响力专利数量之比仅为 1∶12，这导致训练模型的样本非常不平衡，模型将无法充分学习高技术影响力的专利特征。因此，本文在按照7∶3 划分训练集、测试集后，在训练集中复用正样本数据，使得用来构建模型的训练集中正负样本相对均衡。

笔者基于 python 和 scikit-learn 构建 DNN 以及逻辑回归（logistic regression，LR）、随即森林（random forest，RF）、支持向量机（support vector machines，SVM）模型。隐藏层数和每层神经元个数是 DNN 模型的两个关键参数，隐藏层及神经元数量过少会导致模型欠拟合，过多则会导致模型过拟合，均无法得到理想的预测效果。笔者经过多轮实验得到的 DNN 最佳模型由 3 个隐藏层构成，每个隐藏层包含 512 个神经元，每个节点均使用 relu 激活函数，优化器为 Adam，L2 正则化系数为 0.000 1。作为对比试验的 LR 模型、RF 模型和 SVM 三种机器学习模型的参数选择同样经过多轮实验，其中 LR 模型在正则化系数为 1 时效果最好，RF 在包含 7 个深度为 30 的决策树时性能最佳，SVM 模型在惩罚系数为 1，核函数系数为 0.001 时性能最好。

DNN 模型与 LR、RF 和 SVM 三种对比模型的最佳性能对比如表 3 所示。DNN 模型在各项指标上均远高于 LR、RF 和 SVM 模型，这意味着 DNN 在新兴技术预测这一分类任务中的整体性能更好，其不仅能够更全面地识别出未来潜在的新兴技术，尽可能少地遗漏重要的新兴技术专利，从而避免决策者错失发展新兴技术的

机会，又具有较高的精度，能尽可能避免将非新兴技术误判为新兴技术，造成不必要的资源浪费。实验结果表明，DNN 模型相较 LR、RF、SVM 三种广泛使用的机器学习模型性能更优，能更好地拟合指标与技术影响力之间复杂的非线性关系，更有效地进行新兴技术预测。

表 3　DNN、RF、SVM、LR 模型的性能对比

模型	准确率/%	精确率/%	召回率/%	F1-score/%
DNN	95.82	93.73	98.75	96.17
RF	84.38	82.71	87.85	85.2
SVM	68.57	66.46	73.42	69.77
LR	67.31	70.55	63.43	66.8

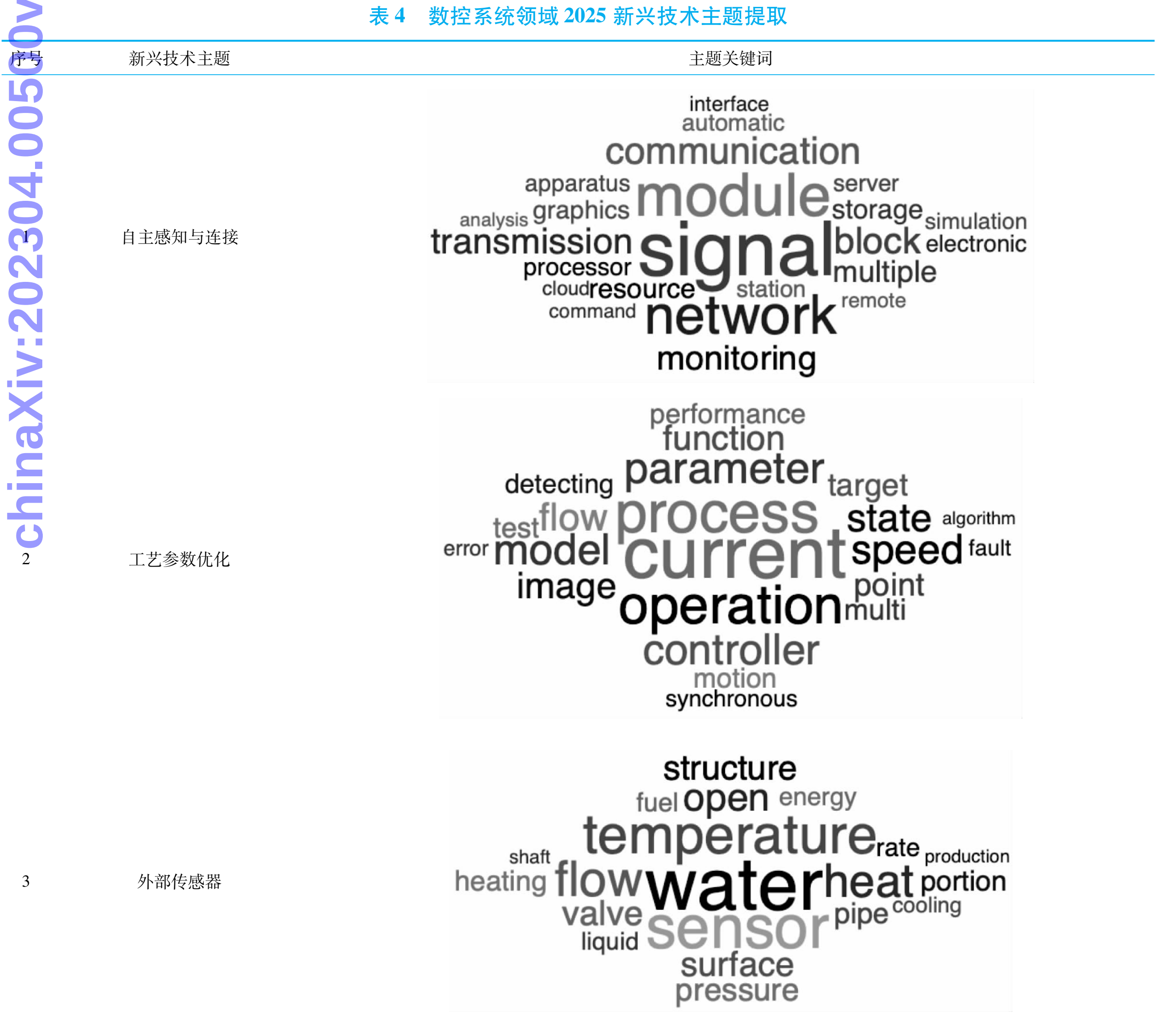
4.3　新兴技术预测与结果分析

4.3.1　新兴技术预测

使用训练完成的 DNN 模型，使用 2016－2020 年的离群专利作为备选新兴技术进行数控系统领域 2025 年的新兴技术预测。计算 2020 年的 15 385 篇离群专利的指标，作为 DNN 模型输入，得到各备选新兴技术 2025 年的技术影响力预测结果。在 15 385 项备选新兴技术中，预计在 2025 年具有高技术影响力的有 348 项，占总离群专利数的 2.26%。

预计在 2025 年将具有高技术影响力的离群专利，即为预测的新兴技术。笔者使用 LDA 模型对新兴技术专利进行主题分析，结合困惑度指标[49]选择主题个数，当主题数为 5 时，困惑度指标下降趋于平缓，且各主题之间区分度较好。对 5 个技术主题提取关键词并绘制词云，结果如表 4 所示：

表 4　数控系统领域 2025 新兴技术主题提取



序号	新兴技术主题	主题关键词
1	自主感知与连接	interface automatic communication apparatus module server analysis graphics storage simulation transmission signal block electronic processor multiple cloud resource station command network remote monitoring
2	工艺参数优化	performance function parameter detecting target test flow process state algorithm error model current speed fault image operation point multi controller motion synchronous
3	外部传感器	structure fuel open energy temperature rate shaft production heating flow water heat portion valve sensor pipe cooling liquid surface pressure

（续表 4）

序号	新兴技术主题	主题关键词
4	误差补偿	protection measuring motor fault mode laser closed phase controller output optical axis drive current loop signal source angle voltage circuit rotating input power feedback shaft fixed supply frequency
5	特种材料加工	concentration membrane substrate ratio polymer material acid zinc reaction binding light layer stream metal hydrocarbon resin components oligonucleotide probe production

根据主题提取结果，数控系统领域 2025 年的新兴技术主要有自主感知与连接、工艺参数优化、外部传感器、误差补偿和特种材料加工 5 个方向：①数控系统作为机床的“大脑”，其自主感知与连接技术是实现智能机床的关键。目前已有的自主感知技术基于“指令域示波器”和“指令域分析方法”[50]。在连接技术方面，美国、德国和中国已先后提出了数控机床互联通信协议，实现制造过程中的信息流传输[51]。国内外企业也已相继推出了数控系统云服务平台，虽然当前这些平台主要停留在技术层面上，但已呈现出应用到智能机床上的潜力与趋势[51]。②在数控加工中工艺参数的优化至关重要，它们影响着零件的加工质量、效率、机床和刀具等制造资源的寿命等。相较传统基于切削稳定性等的建模[52]，基于大数据的建模结合神经网络等人工智能算法，对进给速度、主轴功率等参数进行调整，优化加工工艺[51]。③“互联网 + 传感器”是互联网 + 机床的典型特征，外部传感器加强了对机床加工状态的感知能力[51]，数控系统通过采集机床的温度、压力等数据，并对采集的数据进行分析与处理，实现机床加工过程的自适应控制。④误差补偿是数控系统提供加工质量保障和提升的重要功能，包括热误差补偿、空间几何误差补偿等[51]。数控系统通过机床各部位传感器的反馈数据，结合深度学习等模型[53]进行预测误差并进行补偿，实现全闭环控制。⑤随着航空航天、汽车、医疗等领域对具有硬、脆、热敏、耐腐蚀等性能的特种材料制品的需求日益增长，结合激光加热、电化学等[54]方式的新型材料成型加工技术技术亟待发展，基于增材制造的仿生新材料合成方法也极具潜力[55]。

4.3.2 预测结果分析

用于预测的专利指标可以看作新兴技术的早期特征。为了进一步分析各特征在新兴技术形成过程中的作用，笔者对识别出新兴技术与非新兴技术在各个专利指标维度上进行对比，其概率密度分布结果如图 4 所示。其中浅色、深色分别为新兴技术和非新兴技术在各早期特征上的分布情况。

由图 4 可知，预测出的新兴技术在 PK（先验知识量）、TCT（技术生命周期）、TS（技术范围）、CS（商业范围）、INV（发明人数量）以及 TKH（专利权人能力）6 个指标的分布上与非新兴技术有明显差异。技术融合是新兴技术的重要产生方式，因此从属于多个技术类别，具有较大技术范围的专利更有潜力成为未来的新兴技术。同时，领域内外具有较高能力的专利权人合作，以及多企业机构合作能进一步提升技术创新质量。另外，具有更高的商业价值将提升专利投入实际应用的可能性，进而促进该技术的发展。新兴技术具有更多的先验知识并且基于更早的引用文献，意味着发展新兴技术不能依靠突发奇想，而需要更深入的领域调研，随着技术的发展，一些早期的技术难题或许可以得到解决。新兴技术在 PCID（独立权利要求）、PCD（从属权利要求）两个指标上也略倾向于高于非新兴技术，表

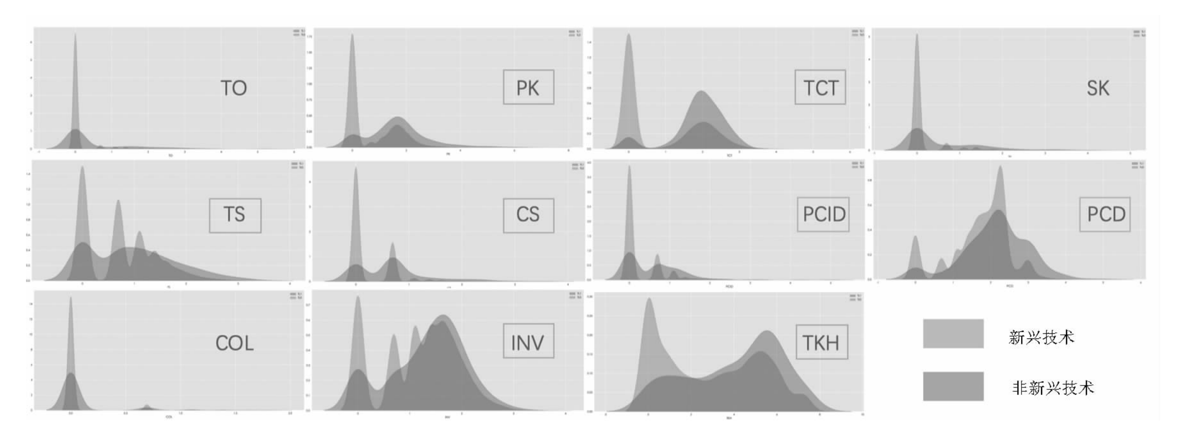


图 4　新兴技术与非新兴技术的早期特征概率密度展示

明由于新兴技术的创新性，其要求的权利保护往往更多。

本实证案例预测了数控系统领域未来具有潜力的新兴技术方向，并分析了新兴技术形成的关键早期特征，验证了笔者提出方法的有效性，对数控系统领域新兴技术的发展与战略布局具有指导意义。

5　研究结论

笔者从离群专利的视角出发采用深度学习方法识别和预测信息技术，基于专利文本相似度筛选出离群专利，进而构建离群专利指标和未来技术影响力之间的关系模型，通过识别领域当前的离群专利预测未来的新兴技术。主要研究结论如下：①根据新兴技术的早期特征，基于离群专利视角识别新兴技术是及时有效的，在数控系统领域的案例研究验证了这一观点；②采用了一种基于 BERT 预训练模型的离群专利识别方法，通过计算专利的文本相似度构建相似度网络以获取与原有技术轨道研究主题差异较大的离群专利；③使用 DNN 构建离群专利指标与技术影响力之间的关系模型，对未来具有更高技术影响力的离群专利进行识别，从而预测潜在的新兴技术，相较 LR、RF、SVM 等机器学习方法效果更好。

笔者从离群专利的视角出发应用深度学习方法进行新兴技术识别，是现有新兴技术预测方法研究视角的重要补充。在离群专利识别过程中，本研究充分利用了专利的文本信息，实现了更有效的相似度计算方法。在新兴技术预测中，拟合了高维专利指标与技术影响力之间复杂的非线性关系，并初步分析了各专利指标在新兴技术形成过程中的作用。本方法时间成本低、适用性广，对任意选定的技术领域均可使用并快速定位潜在的新兴技术。此外，本文识别了数控系统领域的潜在新兴技术，对领域内企业、政府部门的战略布局与规划具有较大的决策支持价值。

本文的方法依然有一定的局限性。首先，笔者仅使用了专利数据描述了新兴技术的早期特征以及技术影响力，主要针对技术驱动的新兴技术预测，未来还可以引入社会影响力、商业效益等指标从而实现更加系统的预测。其次，由于结构较复杂，深度学习模型虽然可以准确拟合出复杂指标与结果之间的关系，尚且无法深入挖掘各指标对结果的影响机制。未来研究可以在关系模型的基础上，进一步发掘其中的因果机制，加强新兴技术预测结果的理论意义。

参考文献：

[1] 薛澜，周源，李应博. 战略性新兴产业创新规律与产业政策研究[M]. 北京：科学出版社，2015：44－56.
[2] VALLE S，VÁZQUEZ-BUSTELO D. Concurrent engineering performance：incremental versus radical innovation[J]. International journal of production economics，2009，119(1)：136－148.
[3] 付玉秀，张洪石. 突破性创新：概念界定与比较[J]. 数量经济技术经济研究，2004，21(3)：73－83.
[4] NOH H，SONG Y-K，LEE S. Identifying emerging core technologies for the future：case study of patents published by leading telecommunication organizations[J]. Telecommunications policy，2016，40(10/11)：956－970.
[5] 李贺，袁翠敏，解梦凡. 专利文献中的睡美人现象分析与研究[J]. 图书情报工作，2019，63(6)：64－74.
[6] 李静海，许光文，杨励丹，等. 一种抑制氮氧化物的无烟燃煤方法及燃煤炉：CN 95102081[P]. 1998－05－20.
[7] SUGITANI H，MATSUDA H，IKEDA M. Liquid jet recording head：US 06/394787 [P]. 1985－12－10.
[8] ROTOLO D，HICKS D，MARTIN B R. What is an emerging technology? [J]. Research policy，2015，44(10)：1827－1843.
[9] 张国胜. 技术变革，范式转换与战略性新兴产业发展：一个演化经济学视角的研究[J]. 产业经济研究，2012(6)：26－32.
[10] AHARONSON B S，SCHILLING M A. Mapping the technological landscape：measuring technology distance，technological footprints，and techny evolution[J]. Research policy，2016，45(1)：81－96.

[11] YOON J, KIM K. Detecting signals of new technological opportunities using semantic patent analysis and outlier detection[J]. Entometrics, 2012, 90(2): 445 - 461.
[12] SONG K, KIM K, LEE S. Identifying promising technologies using patents: a retrospective feature analysis and a prospective needs analysis on outlier patents[J]. Technological forecasting and social change, 2018, 128: 118 - 132.
[13] 罗素平, 寇翠翠, 金金, 等. 基于离群专利的颠覆性技术预测——以中药专利为例[J]. 情报理论与实践, 2019, 42(7): 165 - 170.
[14] ZHOU Y, DONG F, LIU Y, et al. Forecasting emerging technologies using data augmentation and deep learning[J]. Scientometrics, 2020, 123(1): 1 - 29.
[15] CHO Y Y, JEONG G H, KIM S H. A Delphi technology forecasting approach using a semi-Markov concept[J]. Technological forecasting and social change, 1991, 40(3): 273 - 287.
[16] LEE S, KIM W, KIM Y M, et al. The prioritization and verification of IT emerging technologies using an analytic hierarchy process and cluster analysis[J]. Technological forecasting and social change, 2014, 87: 292 - 304.
[17] GEUM Y, LEE S, YOON B, et al. Identifying and evaluating strategic partners for collaborative R&D: index-based approach using patents and publications[J]. Technovation, 2013, 33(6/7): 211 - 224.
[18] SONG B, SEOL H, PARK Y. A patent portfolio-based approach for assessing potential R&D partners: an application of the Shapley value[J]. Technological forecasting and social change, 2016, 103: 156 - 165.
[19] LANJOUW J O, SCHANKERMAN M. Stylized facts of patent litigation: value, scope and ownership[J]. National bureau of economic research, 1997:w6297.
[20] STERNITZKE C, BARTKOWSKI A, SCHRAMM R. Visualizing patent statistics by means of social network analysis tools[J]. World patent information, 2008, 30(2): 115 - 131.
[21] MEYER M. Does science push technology? Patents citing scientific literature[J]. Research policy, 2000, 29(3): 409 - 434.
[22] NARIN F, NOMA E, PERRY R. Patents as indicators of corporate technological strength[J]. Research policy, 1987, 16(2): 143 - 155.
[23] LANJOUW J O, PAKES A, PUTNAM J. How to count patents and value intellectual property: the uses of patent renewal and application data[J]. The journal of industrial economics, 1998, 46(4): 405 - 432.
[24] ARISTODEMOU L, TIETZE F. The state-of-the-art on intellectual property analytics (IPA): a literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data[J]. World patent information, 2018, 55: 37 - 51.
[25] KONG D, ZHOU Y, LIU Y, et al. Using the data mining method to assess the innovation gap: a case of industrial robotics in a catching-up country[J]. Technological forecasting and social change, 2017, 119: 80 - 97.
[26] 周源, 董放, 刘宇飞. 融合新兴领域知识融合过程研究——以生物信息领域为例[J]. 图书情报工作, 2019, 63(8): 127 - 134.
[27] HASSAN S-U, IMRAN M, IQBAL S, et al. Deep context of citations using machine-learning models in scholarly full-text articles[J]. Scientometrics, 2018, 117(3): 1645 - 1662.
[28] ZHOU Y, DONG F, LIU Y, et al. Forecasting emerging technologies using data augmentation and deep learning[J]. Scientometrics, 2020, 123(1): 1 - 29.
[29] GEUM Y, KIM C, LEE S, et al. Technological convergence of IT and BT: evidence from patent analysis[J]. Etri journal, 2012, 34(3): 439 - 449.
[30] KIM G, BAE J. A novel approach to forecast promising technology through patent analysis[J]. Technological forecasting and social change, 2017, 117: 228 - 237.
[31] ZHOU Y, LIN H, LIU Y, et al. A novel method to identify emerging technologies using a semi-supervised topic clustering model: a case of 3D printing industry[J]. Scientometrics, 2019, 120(1): 167 - 185.
[32] 侯剑华, 朱晓清. 基于专利的技术预测评价指标体系及其实证研究[J]. 图书情报工作, 2014, 58(18): 77 - 82.
[33] LEE C, KWON O, KIM M, et al. Early identification of emerging technologies: a machine learning approach using multiple patent indicators[J]. Technological forecasting and social change, 2018, 127: 291 - 303.
[34] 曹艺文, 许海云, 武华维, 等. 基于引文曲线拟合的新兴技术主题的突破性预测——以干细胞领域为例[J]. 图书情报工作, 2020, 64(5): 100 - 113.
[35] ZHANG Y, LU J, LIU F, et al. Does deep learning help topic extraction? a kernel k-means clustering method with word embedding[J]. Journal of informetrics, 2018, 12(4): 1099 - 1117.
[36] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2021 - 05 - 31]. https://arxiv. org/pdf/1810. 04805. pdf.
[37] YANG W, ZHANG H, LIN J. Simple applications of BERT for ad hoc document retrieval[EB/OL]. [2021 - 05 - 31]. https://arxiv. org/abs/1903. 10972. pdf.
[38] BESSEN J. The value of US patents by owner and patent characteristics[J]. Research policy, 2008, 37(5): 932 - 945.
[39] FERN? NDEZ - RIBAS A. International patent strategies of small and large firms: an empirical study of nanotechnology[J]. Review of policy research, 2010, 27(4): 457 - 473.
[40] HAUPT R, KLOYER M, LANGE M. Patent indicators for the technology life cycle development[J]. Research policy, 2007, 36(3): 387 - 398.
[41] BIERLY P, CHAKRABARTI A. Determinants of technology cycle time in the US pharmaceutical industry[J]. R&D management, 1996, 26(2): 115 - 126.
[42] KAYAL A A, WATERS R C. An empirical evaluation of the tech-

nology cycle time indicator as a measure of the pace of technological progress in superconductor technology [J]. IEEE transactions on engineering management, 1999, 46(2): 127－131.

[43] COZZENS S, GATCHAIR S, KANG J, et al. Emerging technologies: quantitative identification and measurement [J]. Technology analysis & strategic management, 2010, 22(3): 361－376.

[44] DAY G S, SCHOEMAKER P J. Avoiding the pitfalls of emerging technologies [J]. California management review, 2000, 42(2): 8－33.

[45] GUELLEC D, DE LA POTTERIE B V P. Applications, grants and the value of patent [J]. Economics letters, 2000, 69(1): 109－114.

[46] MA Z, LEE Y. Patent application and technological collaboration in inventive activities: 1980－2005[J]. Technovation, 2008, 28(6): 379－390.

[47] MEYER M. Are patenting scientists the better scholars? An exploratory comparison of inventor-authors with their non-inventing peers in nano-science and technology [J]. Research policy, 2006, 35(10): 1646－1662.

[48] HARHOFF D, NARIN F, SCHERER F M, et al. Citation frequency and the value of patented inventions [J]. Review of economics and statistics, 1999, 81(3): 511－515.

[49] 董放，刘宇飞，周源. 基于 LDA-SVM 论文摘要多分类新兴技术预测[J]. 情报杂志，2017(7): 40－45.

[50] CHEN J, YANG J, ZHOU H, et al. CPS modeling of CNC machine tool work processes using an instruction-domain based approach[J]. Engineering, 2015, 1(2): 247－260.

[51] CHEN J, HU P, ZHOU H, et al. Toward intelligent machine tool [J]. Engineering, 2019, 5(4): 679－690.

[52] SAFFAR R J, RAZFAR M. Simulation of end milling operation for predicting cutting forces to minimize tool deflection by genetic algorithm[J]. Machining science and technology, 2010, 14(1): 81－101.

[53] LI Z, WANG Y, WANG K. A data-driven method based on deep belief networks for backlash error prediction in machining centers[J]. Journal of intelligent manufacturing, 2020, 31(7): 1693－1705.

[54] 陈小丽. 硬脆性材料复合加工技术综述[J]. 航空发动机，2010(3): 57－60.

[55] POLLINI B, PIETRONI L, MASCITTI J, et al. Towards a new material culture. Bio-inspired design, parametric modeling, material design, digital manufacture[C]// Design in the digital age technology, Nature, Culture. Milano: Politecnica University Press, 2020: 208－212.

作者贡献说明：

孔德婧：负责框架设计，论文修改及撰写指导；

董放：负责观点提炼，实验指导；

陈子婧：负责实验设计与论文实验，论文撰写；

刘宇涵：负责论文实验；

周源：负责论文撰写指导。

Prediction of Emerging Technologies from the Perspective of Outlier Patents ——Based on Bert Model and Deep Neural Networks

Kong Dejing[1] Dong Fang[2] Chen Zijing[3] Liu Yuhan[3] Zhou Yuan[2]

[1] School of Modern Post, Beijing University of Posts and Telecommunications, Beijing 100876

[2] School of Public Policy and Management, Tsinghua University, Beijing 100084

[3] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074

Abstract: [**Purpose/significance**] Due to the advanced nature of emerging technologies, they are often marginalized at the initial stage of formation. Most of present researches forecast emerging technologies by analyzing the mainstream technology development path, which would neglect some research that disrupts existing technology routes. By analyzing outlier patents that are less similar to the mainstream technologies in the field, it can identify and forecast the future emerging technologies more effectively. [**Method/process**] This paper presented an outlier patent identification and emerging technology prediction method based on deep learning. Firstly, the Bert pre-trained model was used to construct the similarity network based on texts of patents and outlier patents identification. The relationship model between outlier patent indicators and technical influence was then built based on DNN model, thus realizing the fast and accurate emerging technology prediction using large-scale outlier patents. Finally, an empirical analysis was conducted in the field of numerical control system with all patents applied in the last ten years obtained from DI database. [**Result/conclusion**] The result of empirical analysis in the field of numerical control system not only verifies the validity of the model, but also has important guiding significance to the formulation of national technology development policy and the technology layout of enterprises in related fields.

Keywords: emerging technologies deep learning outlier patents numerical control system